Blumea 54, 2009: 16–17
www.ingentaconnect.com/content/nhn/blumea

RESEARCH ARTICLE

doi:10.3767/000651909X474041

Bioinformatics interoperability: all together now !

B. Meganck[1], P. Mergen[1], D. Meirte[1]

Key words

biodiversity informatics
interoperability
Web 2.0

Abstract The following text presents some personal ideas about the way (bio)informatics[2] is heading, along with some examples of how our institution – the Royal Museum for Central Africa (RMCA) – is gearing up for these new times ahead. It tries to find the important trends amongst the buzzwords, and to demonstrate how these will benefit the biological and scientific community. This text is based on a presentation given at the 7th Flora Malesiana Symposium in Leiden, The Netherlands, June 2007.

Published on 30 October 2009

BIOINFORMATICS AS INFORMATICS FOR BIOLOGISTS

For some years now, computers have simplified biologists' lives. Or have they? They surely brought a host of new complications, frustration and headaches. And new ways to lose your data. But if you know how to handle them, and if everything goes just right, they can help with some of the tedious sorting and processing necessary for extracting the information from the samples. And if you are really good you can, by putting in long hours and much study, use them to produce nice graphs or maps, and put these on the web. But as it is, bioinformatics is still mostly a thing for geeks. The programs can be very powerful, but they all require (much) experience to be handled properly.

First, there are various and difficult interfaces, each one with a learning curve. And if you switch to another piece of software, you can start all over again. Who has time for that?

Secondly, one tool in itself is never enough. If you want to sort and clean up your data, you need something like a spreadsheet. If you want to store your data after that you need a database. If you want to make maps you have to find a GIS system. If you want these maps on the web you have to find and set up a WebMapService (WMS) and write your HTML interface. There is not just one software package you are required to master, but five or six. So the headaches multiply.

Last but not least, there is a bewildering jungle of file formats, making every exchange a challenge, and draining your last bit of desire ever to touch a computer again. Why won't this shapefile go straight into the database? Shall I save my maps as .jpg, as .png, as .pdf, as .svg? Which software will open this .mif file, and why am I writing three files when storing a single map?

DEVELOPMENTS IN BIOINFORMATICS

As it has always been in computer-land, part of the confusion is due to 'hypes' giving us *Wikis*, *Youtube*, *MySpace*, and *Flickr;* and 'buzzwords' such as *Blogs*, *OGC*, *XML*, and *cross-platform.* In addition, there are *open standards*, and *Google Earth*, and lots more.

Looking at it all from a little distance, can we perhaps get some hints about where informatics is headed? For a start, none of these technologies present a major technological breakthrough. In fact, all these examples use mostly technologies that have been around since the 1970s or 1980s: we've seen satellite images in our TV weather forecast for ages, 3D image rendering has been around for a long while, GIS and mapping systems are already mainstream as well. It is not the technology part that makes the buzz. What's new is how these technologies are used, and by whom.

Exponential increases in computing power and Internet penetration have taken these applications – once the monopoly of big institutions – and put them into the hands of mere mortals like you and me. Even better: we're not just allowed to use these goodies, we're even asked to provide the input data! So there is a shift towards community-driven sites, where ordinary people act both as data providers and consumers. This was something unheard of until recently, and this new order is broadly encompassed by the über-buzzword: *Web 2.0*.

Web 2.0 boils down to three Key words:
- standards
- cooperation
- human interface

Existing tools and techniques are thus strung together by standards for data exchange, so that they can cooperate smoothly, while presenting a human-friendly face to the user. Or, alternatively: the isolated technologies of the Web are learning to cooperate, and how to interact with these strange, non-deterministic entities: human beings. As a result, 'informatics for geeks' is slowly turning into 'informatics for all'.

STANDARDS AT THE ROYAL MUSEUM FOR CENTRAL AFRICA

The basic key word in this development is standards: agreements about the way to cooperate.

Good standardisation is what makes you able to plug in your razor in the sockets of any country you happen to be in. Imperfect standardisation is what forces you to lug around clumsy adapters.

In bioinformatics, there are standards as well: common languages to define the information we are sharing. For example, the *ABCD* and *Darwincore* standards offer frameworks for describing biological data (specimens and collections) in a number of pre-set fields. In both standards, you would easily recognise the fields containing the obvious information (specimen number,

[1] Biodiversity Information and Cybertaxonomy Group, African Zoology Department, Royal Museum for Central Africa, Tervuren, Belgium.

[2] In literature, the term 'bioinformatics' is mostly restricted to the context of molecular biology and software. In these paragraphs I use it less strictly for any software solution useful for biologists.

scientific name, collector) one would like to have about a specimen. A tool that 'speaks' ABCD can connect to all those who offer their data in this format, and things will just work. If you ask for the family name of specimen *RMCA0123*, you will get it. If you want the collector and the name it is already there.

Other standards exist, adapted to other purposes: *GML* for example, is roughly similar to ABCD, but aimed squarely at exchanging geographical information. So GML is not talking about field numbers and taxa, but about coordinates, geographical strata, and rock samples. WMS and WFS are the standards for putting maps on the web. Once you've braved the tsunami of acronyms, all these standards do a good job, and they're not prohibitively difficult to use.

The Royal Museum for Central Africa (RMCA) has put these standards (and others) to good use, and has benefited from it. A real-life example will make this a bit more clear:

Our Herpetology department recently participated in the HerpNet project, an on-line data portal of amphibians and reptiles (http://www.herpnet.org). HerpNet asked for communication following the Darwin Core protocol, and so we installed a DarwinCore server on top of our database. Now our records can be queried by everyone, anywhere in the world through the HerpNet portal.

But with the tedious work of checking the specimens done, we were anxious to do even more with our data. So next to the DarwinCore server, we installed an ABCD server – in order to talk to the GBIF data portal. With half a day's work, the same records from the same database are now accessible through GBIF as well – giving them much more visibility, since GBIF is very well known and widely used in biodiversity informatics (http://www.gbif.org).

Furthermore, nothing would stop us from installing a WMS service on that database as well. That would take a few days of our time, but it would enable anyone to display our data points on a (screen) map in Google Earth, in NASA Worldwind, and elsewhere.

Our SYNTHESYS NA_D 3.7 "Itinerary" project (http://synthesys.africamuseum.be/home.html), retracing expedition pathways from the gathering places of the specimens, already uses a WMS service.

A DEMO FOR A SCIENTIFIC WORKFLOW

The RMCA has participated in a GBIF/TDWG-organised workshop aimed at demonstrating the ease of use of standard-compliant tools (see http://wiki.tdwg.org/twiki/bin/view/Geospatial/InteroperabilityWorkshop1). Within a week, seven programmers from different countries produced a scientific workflow, chaining together existing web services behind a user-friendly interface – in line with the *Web 2.0* paradigm.

Our tool offers a consecutive series of screens, where the user can:

- search on a valid name in Catalogue of Life;
- search occurrences for that name in GBIF, or other online providers;
- select appropriate environmental layers (average temperature, rainfall, …);
- run a modelling program with the occurrence points and the environmental parameters;
- display the results on a map

So, in 5 or 6 easy steps he or she can complete a useful scientific survey, without any in-depth knowledge about the (intricate) software behind the interface.

Brushing any false modesty aside, we were really pleased with this result and it should bode well for any scientist, too. It demonstrated that online tools that are fit for daily use by biologists without any technical knowledge, are just around the corner. It demonstrated that the *Web 2.0* philosophy does, in fact, work in everyday situations. And, best of all, it shows that it is realistic to expect that bioinformatics may become what it should be: a tool that allows scientists once again to focus on their core business – Biology

REFERENCES

Websites:
http://www.gbif.org
http://wiki.tdwg.org/twiki/bin/view/Geospatial/InteroperabilityWorkshop1
http://www.herpnet.org
http://synthesys.africamuseum.be/home.html

Appendix Glossary for non-geeks.

blog — a blog or 'weblog' is an online diary you keep, with regular updates on your thoughts and adventures, sharing them with the Internet community.

cross-platform — software that you can use on any computer system (platform): Linux, Mac (Apple), Windows …

Flickr — an online 'kiosk' where anyone can post his personal (digital) photos. So after a trip you can dump your holiday pictures for all your friends (and everyone else) to see.

GIS — Geographical Information System. In a word: maps on your computer. These maps are build from layers, each representing a geographic/geologic entity, for example a layer with rivers and lakes, a layer with vegetation, a layer with rock types, a layer with roads and cities, and so on. You can enable or disable any of these layers, and thus explore the region and your data points. Are all data points located near rivers? Which points follow the roads?

Google Earth — A very nice tool/toy developed by Keyhole software, and taken over by Google. It's a virtual globe that appears on your screen, and that you can manipulate (rotate, zoom in). The earth surface is covered with satellite images, some of which are very detailed (so you can see individual houses and cars). People can add points of interest, and upload pictures (e.g. the Eiffel tower in Paris). NASA Worldwind is pretty much the same thing, only with more focus on openness, scientific correctness and data attribution.

HTML — Hypertext markup language. The language used for writing web pages. Basically, it is text with some instructions for layout on the page (font type, spacing, alignment, …).

KML — Keyhole Markup Language. The XML language used by Keyhole software for its earth viewer. Now the earth viewer has become Google Earth, but KML has not changed its name. You can display KML files directly in Google Earth.

MySpace — a social networking website where you can present yourself, write your blogs, link to your friend's homepages, …

NASA Worldwind — see Google Earth.

OGC — Open Geospatial Consortium: a group of interested parties setting standards for geospatial applications (maps, GIS systems). OGC developed, amongst many others, the WMS and WFS standards.

WFS — a Web Feature Service; an addition to WMS. The Web Feature Service enables you to query the points displayed on an online map. So, things get interactive: you can get the name (or number) of the point, and any other information attached to it.

wiki — a wiki is a collaborative, web-based documentation platform. Anyone can write or alter texts (with or without review) and thus add information. Wikipedia is one of the best-known applications of this technique.

WMS — a Web Map Service. An online service that provides you with maps. You specify the region you want, the size of the map, the colours to be used … and you receive an image file with your map.

XML — eXtensible Markup Language. A format for storing data in text files. The data fields are enclosed within clear and explicit 'tags'. So an author's name would be something like:
<author_name> Linnaeus </author_name>.
That's clear and even human-understandable. Best of all, XML files can be used on any computer system (Unix, Apple MacOSX, Windows, Linux, Solaris, BeOS, ...), so they make for easy cooperation. KML and GML are examples of XML-type formats.

YouTube — a video sharing website: anyone can upload a video if they feel others will be interested.